Text Variation Explorer

Towards interactive visualization tools for corpus linguistics*

Harri Siirtola[1], Tanja Säily[2], Terttu Nevalainen[2] and Kari-Jouko Räihä[1]
[1]University of Tampere / [2]University of Helsinki

This is a post-print version of the following article: Siirtola, H., Säily, T., Nevalainen, T. & Räihä, K.-J. 2014. "Text Variation Explorer: Towards interactive visualization tools for corpus linguistics". *International Journal of Corpus Linguistics,* 19 (3), 417–429 (doi:10.1075/ijcl.19.3.05sii).

This paper reviews the gap between current methods of text visualization and the needs of corpus-linguistic research, and introduces a tool that takes a step towards bridging that gap. Current text visualization methods tend to treat the problem as a data-encoding issue only, and do not strive for interactive, tightly coupled representations of text that would foster discovery. The paper argues that such visualizations should always be *linked* for effortless movement between the text and its visualization, and that the visualization controls should provide *continuous* and *immediate feedback* to facilitate exploration. We introduce a tool, *Text Variation Explorer* (TVE), to demonstrate the aforementioned requirements. TVE allows visual and interactive examining of the behaviour of linguistic parameters affected by text window size and overlap, and in addition, performs interactive principal component analysis based on a user-given set of words.

Keywords: text visualization, interactive exploration

1. Introduction

Information visualization is a cross-disciplinary field that aims to amplify human cognition with external tools that make information acquisition or reasoning easier (Spence 2007). Often these external tools are visual as more information is acquired through vision than through all the other senses combined (Ware 2004: 2). The strength of the information visualization approach is that it often suggests interesting aspects of data that are difficult to realize using statistical methods alone. If the user is allowed to interact with the data, then the chances for exploratory findings are even better. It has been suggested that interaction and inquiry are in fact inextricable (Pike et al. 2009).

Text visualization is a challenging area (see Hearst 2009: Ch. 11 or Card et al. 1999: Ch. 5 for a review). There are few text visualization tools that are interactive, and provide continuous and immediate feedback. The majority of text visualization tools, such as *DocuScope* (Kaufer et al. 2006) and *WordSift* (Hakuta 2011), mainly support stepped interaction (i.e. a mouse click causes movement in discrete information space, Spence 2007: 5), which is the prevalent mode of interaction in web-based tools.

In this paper we assess the benefits and ramifications of an information visualization approach to corpus linguists, and discuss some of our linguistically motivated visualizations. Although general-purpose visualization techniques provide a good starting point, techniques that dig deeper into the structure of the texts included in a corpus, and work bottom-up from the texts, are needed to gain insight into linguistic

variation and change. For instance, the open-source tool *Mondrian* (Theus 2011) is ideal for quickly formulating hypotheses about data, and even for verifying them in some cases (Theus & Urbanek 2008; for a linguistic use, see Siirtola et al. 2011). However, the interactive graphs produced by *Mondrian* are missing one essential thing: the connection to the text itself.

In the next section we discuss the requirements that corpus linguistics sets for information visualization methods. We then present the linguistically motivated visualization tool that we have developed, and finally, discuss how linguists might benefit from the information visualization approach in their work.

2. Desiderata for corpus-linguistic visualization tools

Text is a challenging data type. Unlike other data types, text does not have a fixed meaning. Put in a different order or in a different context, words may assume a different meaning, as may the words spoken or written *by* a certain person or *to* a certain person. The extreme view is that a text corpus does not have meaning or functions at all – just frequencies of word occurrence and co-occurrence (Gries 2009: 11). One thing is clear: it is a challenge to quantify text so as to make the best possible use of computational methods. While the study of language must abstract away some detail to be able to generalize and draw conclusions, corpus linguists wish to make sure that the generalizations hold for actual language use, which means that the analysis should not lose the connection to the text on which it is based.

In our view, corpus linguists need at least three kinds of visual analysis tools:

i. exploratory visualization and analysis tools (which are our focus here);
ii. explanatory tools to make a point; and
iii. tools for statistical, confirmatory analysis.

The distinction between these categories is often misunderstood, and statistical graphics are seen as a sufficient means to make exploratory observations. But there is a fundamental difference between exploratory and presentation graphics. As Theus & Urbanek (2008: 6) point out, "the relation between the number of observers and the number of graphics in use is inverse". In exploration, a huge number of graphics is created for a single observer, and in presentation, a single graphic must serve a huge number of observers. An attempt to serve both purposes with the same graphic is always deemed to be a compromise, and invariably a mediocre one.

Another important distinction between these categories is interaction. Statistical analysis tools rarely support the continuous, direct manipulation style of interaction that is highly valuable for pattern discovery and insight generation. Instead, the stepped mode of interaction is standard in statistical tools.

Shneiderman's (1996: 337) famous visual information seeking mantra, "overview first, zoom and filter, then details on demand" is a principle he finds recurring in his own designs (see Craft & Cairns 2005 for discussion). However, applying the mantra in the context of corpus linguistics leads to immediate problems. In an exploratory visualization, a corpus linguist wishes to see the *connection* to the relevant part of the text at all times, so the *detail* should always be there. Perhaps the corpus-linguistic mantra should read "text first, then text with a visualization".

There are very few exploratory visualization and analysis tools that are linguistically motivated (but see Culy & Lyding 2010, Hilpert 2011), and even fewer that allow rapid exploration of linguistic parameters. We developed our tool in a user-centred manner, starting from an idea conceived by linguists in an iterative process. The next section describes this tool.

3. *Text Variation Explorer*

To consider the benefits of interactive information visualization for corpus linguists, we present a problem of finding an optimal window size for linguistic parameters (Section 3.1), and propose a tool to solve this problem. *Text Variation Explorer* allows visual and interactive examining of the behaviour of linguistic parameters affected by text window size and overlap (Section 3.2), and in addition, performs interactive principal component analysis based on a user-given set of words (Section 3.3).

3.1 Problem description

Corpus compilers typically sample texts using a fixed sample size: the influential Brown family is based on 2,000-word samples, and the London-Lund Corpus of Spoken English has texts consisting of 5,000 words each. As Kilgarriff (2012: 130) notes, a fixed sample length greatly facilitates the statistical comparison of corpora. But since there is no standard sample size, automatic comparisons are not possible. Moreover, different sizes may be needed for computing different text measures. Biber (1988: 238–239), for example, takes the first 400 words of each text (sample) in a corpus to compute the type-token ratio; normalizing texts of different length would skew the results. Biber et al. (2007: 161) identify discourse units in texts by comparing similarity scores based on type frequencies in two adjacent 50-word sliding windows. Keim & Oelke (2007) find that *hapax legomena* stabilize at 1,300 words, and Lijffijt et al. (2012) go as far as to argue that a specific algorithm is needed for determining the optimal window size for each measure. But they also note that different window lengths themselves can show interesting properties of the data. Hence being able to control sample length makes it possible to compare findings across corpora and the measures used.

It can therefore be argued that, besides taking a predefined or computationally determined window size, we could take the exploratory route: vary the window size and observe when a linguistic parameter reveals interesting patterns in the text. Too small a window may make the parameter to fluctuate rapidly, and too large a window may produce no interesting changes at all. An optimal window size would flag a potentially interesting phenomenon in the text. Initially, we experimented with a large number of linguistic parameters – apart from those selected, e.g. Honoré's R, Simpson's index and the proportion of dis legomena – but our tests revealed that they were mostly redundant in terms of flagging interesting changes. We ended up using a set of three parameters: type-token ratio (TTR), the proportion of *hapax legomena* (words appearing once in the text fragment), and the average word length. Table 1 shows these measures for a brief text excerpt with a tiny window size to illustrate the concepts.

Table 1. A brief text excerpt with three different measures

Text fragment with a window size of 9 words, and an overlap of 7 words	**TTR**	**Hapax legomena**	**Average word length**
Baker Street was like an oven, and the glare	1.00	1.00	3.89
was like an oven, and the glare of the	0.89	0.88	3.22
an oven, and the glare of the sunlight upon	0.89	0.88	3.78
and the glare of the sunlight upon the yellow	0.78	0.86	4.11
glare of the sunlight upon the yellow brickwork of	0.78	0.71	4.67
the sunlight upon the yellow brickwork of the house	0.78	0.86	4.78
upon the yellow brickwork of the house across the	0.78	0.86	4.56
yellow brickwork of the house across the road was	0.89	0.88	4.56

A more realistic example would be something like the text of a novel. For instance, the size of James Joyce's *Ulysses* is 266,306 words (if we define a *word* to be anything separated by whitespace or the characters -+/=#%,;:"!?.). If we compute a table like Table 1 with a more reasonable window size of 200 and an overlap of 50, the result will be a 4 × 1,776 table. It is very hard to detect interesting patterns from such a table, or to compare it with a table computed with slightly different parameters. A better approach is to produce a line graph of the measures (Figure 1).

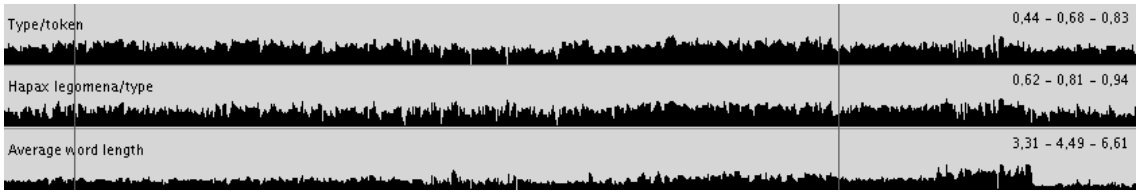

Figure 1. Line graph of measures for James Joyce's *Ulysses* with window and overlap values of 200 and 50 words, respectively

The line graphs for the three measures in Figure 1 have been normalized, i.e. they fill all the available space. The minimum, average, and maximum values for each measure can be read from the right end of the chart. It is fairly apparent from this graph that the end of the novel is somehow different, but the window size is perhaps not optimal to reveal it. Figure 2 shows the line graphs with a window size of 1,325. Now it is apparent that the end of the novel is different in terms of these three measures. The last part of the novel is indeed a soliloquy, written in an experimental stream-of-consciousness style, containing eight run-on sentences without punctuation.

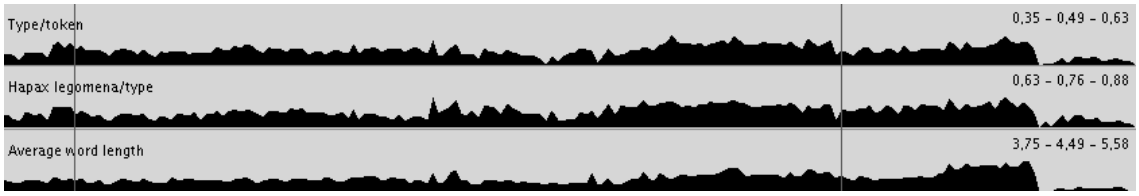

Figure 2. Line graph of measures for James Joyce's *Ulysses* with window and overlap values of 1325 and 50 words, respectively

Figure 2 shows a good deal more than just the divergent style of the novel ending, however. A linguist can use it to explore text passages corresponding to peaks and valleys in parameter values, identifying possible reasons for the variation (cf. Youmans 1991). For instance, in a novel, peaks in the TTR might correspond to narrative sections, while valleys could indicate internal or external dialogue.

However, what is crucial even to the experienced analyst is the chance to see the underlying text once something interesting is observed, or to move quickly between the graph and the text. For this functionality we need an interactive computer application which is described next.

3.2 *Text Variation Explorer*

Text Variation Explorer (TVE, Siirtola 2012) is an application to visually and rapidly seek the most "interesting" fragment size (analysis window) for a given text. Figure 3 shows two elements from TVE's user interface: a text pane where the text to be analysed is pasted, and the controls for the text fragment settings. The Word break field defines a set of characters that cannot appear within a word, and the Word count field shows into how many words the current word break set splits the text. The Window and Overlap sliders set the desired limits, and the Fragment count specifies the number of segments the text is divided into using these sliders. These sliders update the line graphs after every change, and allow rapid exploration of a large number of different window size and overlap settings.

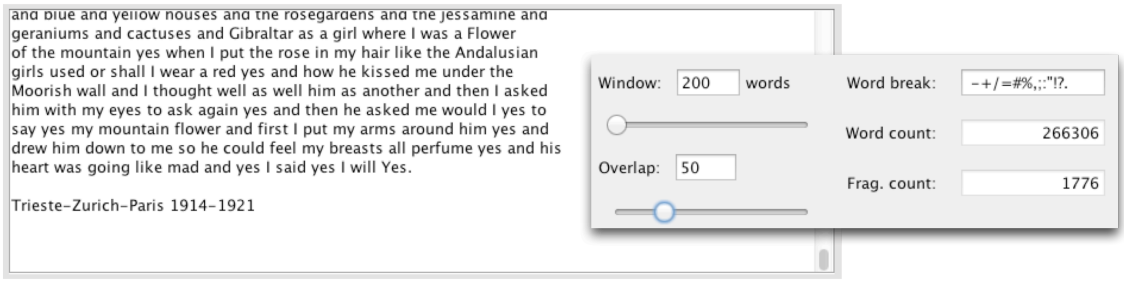

Figure 3. TVE's text pane and window setting controls

TVE implements brushing between the text view and the line graph view of measures. Clicking an interesting-looking point in the line graph will highlight the corresponding fragment in the text pane, and vice versa (Figure 4). This simple functionality is what turns a fairly conventional graphing application into a linguistic visualization tool.

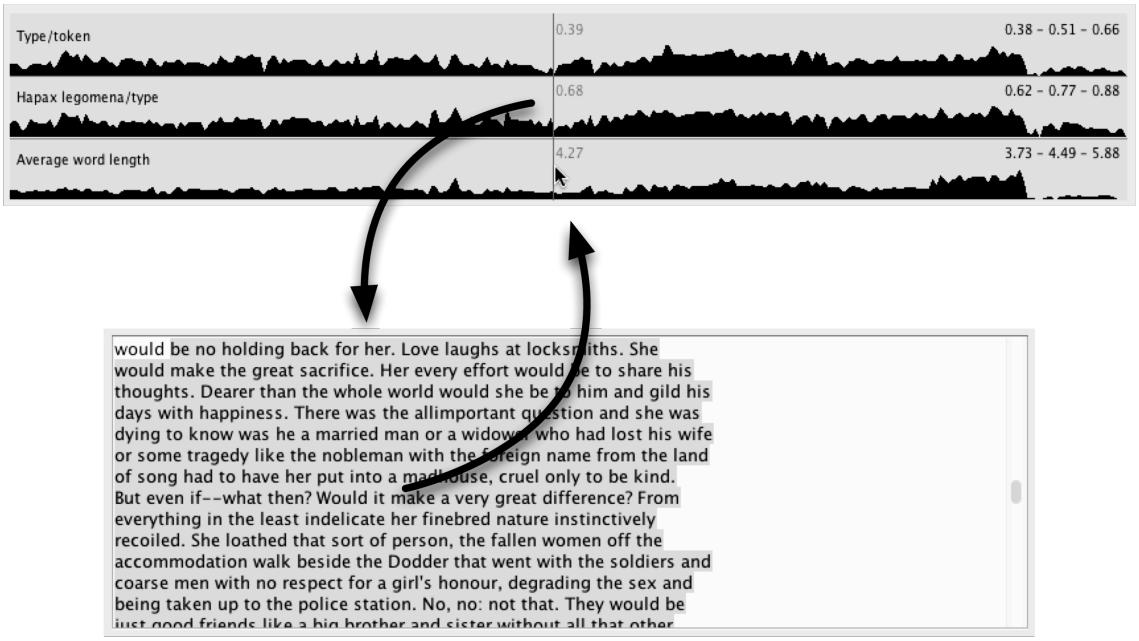

Figure 4. Brushing between views

The line graph view has a zoom facility for situations when the number of text fragments exceeds the number of available screen pixels. The zoom is activated by a mouse press, and remains zoomed-in until the mouse button is released, allowing the user to see every single measure in the line graph.

3.3 Text clustering with TVE

TVE can also cluster the text fragments according to a user-given set of words (Siirtola 2013). TVE performs a principal component analysis (PCA) on the frequencies of the given words in the text fragments. For example, if we enter 55 personal pronouns and have 268 text fragments, TVE will perform PCA on the table of 268 × 55 word frequencies, rotating the 55-dimensional data set in such a way that the first axis (or component) has the largest deviation, and the second axis has the next largest deviation.

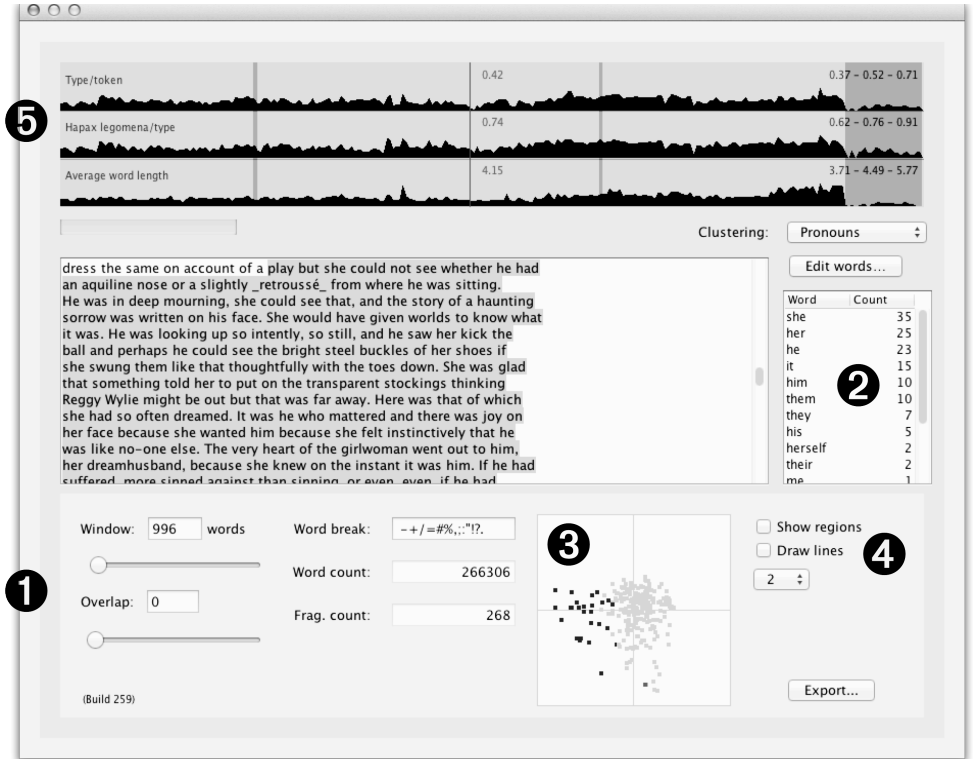

Figure 5. Clustering the text fragments: James Joyce's *Ulysses* clustered according to 55 personal pronouns

Figure 5 shows, again, James Joyce's *Ulysses* in TVE. The 266,306-word novel has been divided into 268 fragments of 996 words (Figure 5, (1)), and clustered according to 55 personal pronouns (Figure 5, (2)). The principal component view in Figure 5 (3) displays each text fragment as a point, and shows the values of the first two principal components for it. If the points in view (3) appear together, it indicates that their distribution of pronouns is similar. The points can also be coloured with a *K*-means algorithm by requesting TVE to force a certain number of groups (Figure 5, (4)). Besides colouring the points, TVE employs the same colour coding in the line graph as well, making it easy to see if similar text fragments are continuous in the text. The adjacency of text fragments can also be seen by requesting TVE to connect the PCA dots ("draw lines", (4)). Another option is to request TVE to show minimal convex hulls that enclose the points in a cluster ("show regions", (4)), which is implemented with Graham's scan algorithm. The PCA view implements three-way brushing between the text and the line graph view, i.e. selecting a text fragment in any view will propagate the selection into other views as well.

The situation shown in Figure 5 suggests that the end of *Ulysses* is different in its use of pronouns than the earlier parts of the novel. The left cluster indicated by the PCA view contains fragments where the pronoun *I* is prevalent. Similarly, the right cluster is dominated by the pronouns *she*, *her*, *him*, and *he*. The last part of the novel is clearly different in this respect as well.

Let us take another example, this time from a linguistic corpus, namely the Brown family mentioned in Section 3.1 above. These are one-million-word corpora of British and American English from the 1960s and the 1990s, with a 1930s extension in the making. They are built to be comparable, each including the same number of texts

from different genres. But how similar are they? In Figure 6, we have pasted both the original Brown corpus (American English from the 1960s) and the Lancaster-Oslo/Bergen corpus (LOB; British English from the 1960s) in TVE. To make it clear where Brown ends and LOB begins, we have inserted the word *dammocmark* between them (DAMMOC was the name of the project where TVE was developed). Wherever this word is inserted, TVE draws a vertical blue line in the line graph view, facilitating comparisons between different corpora and texts.

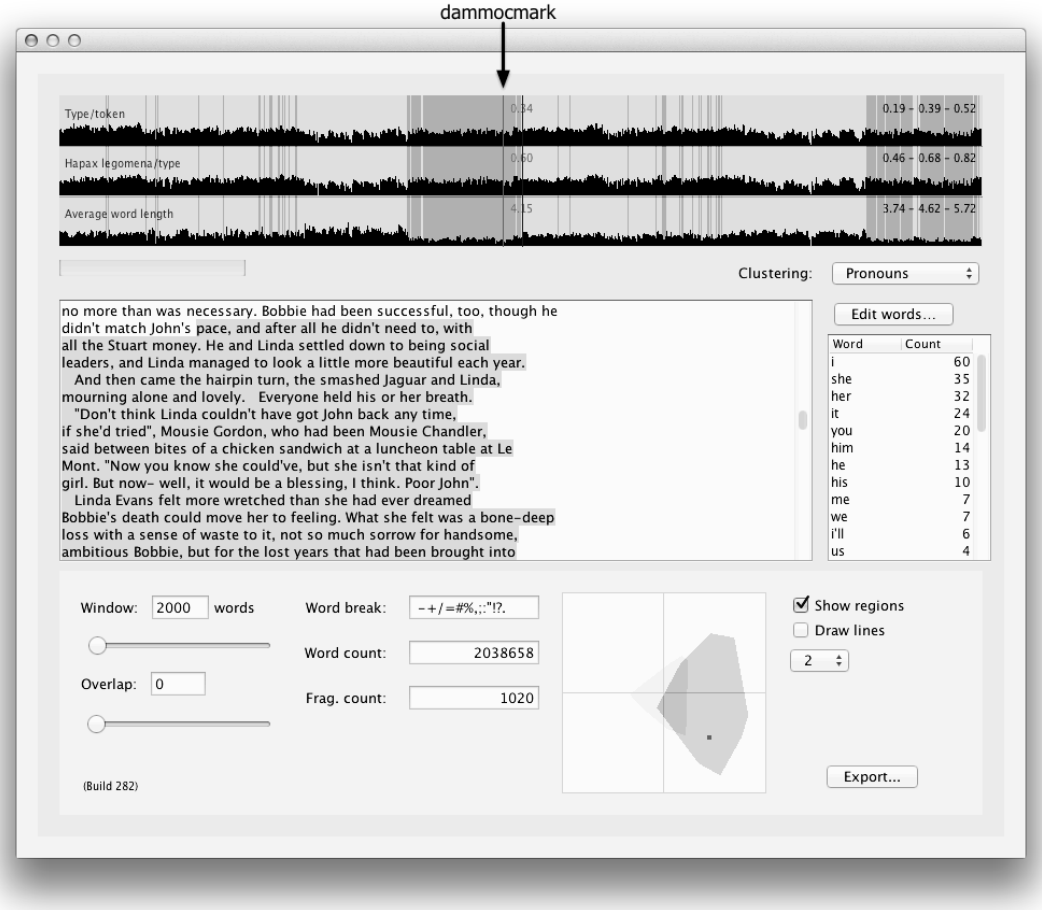

Figure 6. Clustering the text fragments: Brown and LOB clustered according to personal pronouns

Because we know that the Brown family is based on 2,000-word samples, we use a window size of 2,000 words in TVE. To minimize the amount of noise, we set the overlap to zero. In Figure 6, we cluster the text fragments based on a list of personal pronouns, as in the *Ulysses* example. Both of the corpora seem to be divided fairly neatly into two sections. When we click on the different colours in the PCA view to see what kinds of texts appear in the text view, we discover that the two sections seem to represent non-fiction and fiction. Thus, the use of personal pronouns separates fiction from non-fiction, and the British and American corpora seem to use pronouns similarly within each section. Figure 7 shows the result of changing the list of personal pronouns to a list of function words from Binongo (2003). Function words seem to separate

fiction from non-fiction equally well as pronouns did, indicating systematic differences between these domains.

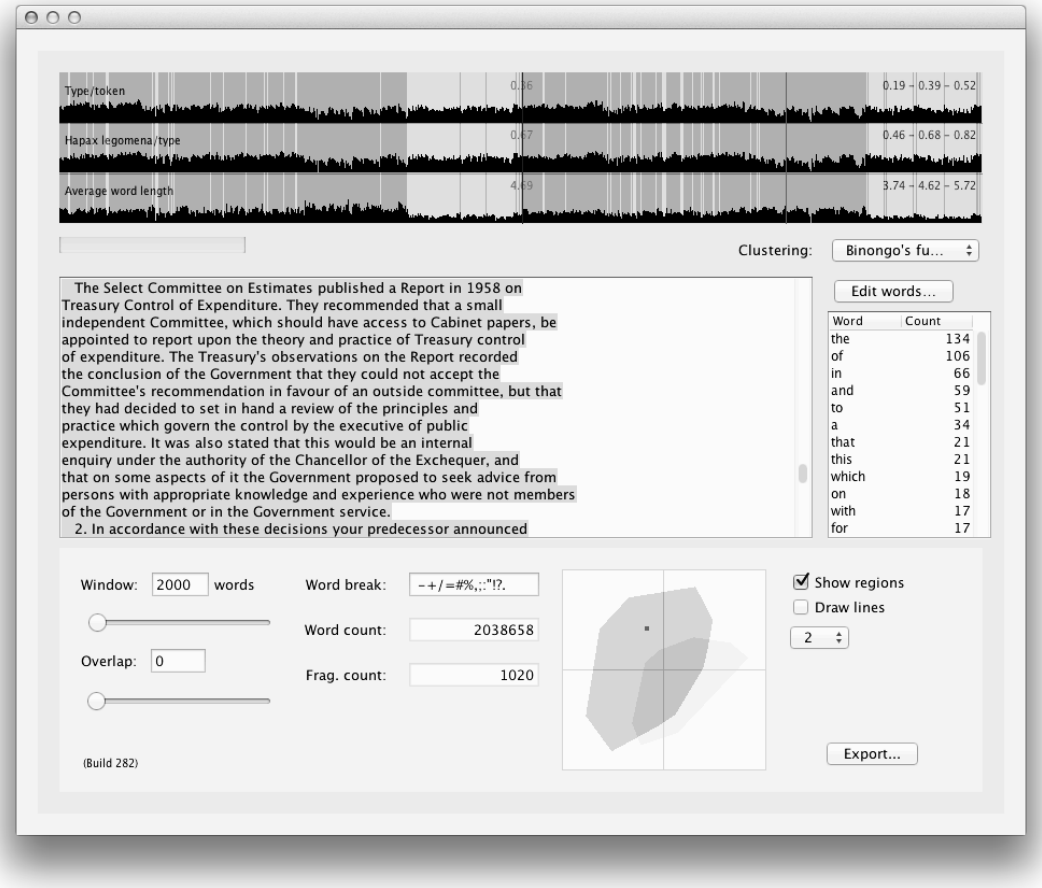

Figure 7. Clustering the text fragments: Brown and LOB clustered according to Binongo's (2003) list of function words

Another set of corpora that could be explored in this manner is the International Corpus of English (ICE). These corpora of English worldwide are meant to be comparable, but as this is a major endeavour involving dozens of researchers in various countries, the compilation conditions have not always been identical. TVE could help users to gain a quick overview of similarities and differences across the corpora, highlighting sections that require more careful analysis.

Text Variation Explorer is intended as a visual exploration tool for partitioning text in ways that look promising to the analyst, to be further inspected with other tools, such as statistical programs. The approach to PCA implemented in TVE is to visualise only the first two principal components, and to use all the available space to plot their two-dimensional values. This invariably leads to different scales on principal components, and may exaggerate differences. The aim is not to give false information, but to make sure that all the differences are detected. TVE is a tool for the discovery of a phenomenon, but not for the verification of it. Therefore, the tool includes an export functionality that writes out the text fragment data in a tab-delimited form for further analysis, e.g. in the statistical system *R*.

4. Discussion and conclusion

We have considered the benefits of interactive information visualization for corpus linguists by presenting a problem of finding an optimal window size for linguistic parameters, and by proposing a tool to solve this problem. We advocate the use of exploratory, highly interactive analysis techniques especially in situations where the goal is not well defined. This approach is rather different from the current state of corpus-linguistic research where simple text concordancers and spreadsheet applications are still the prevailing tools. However, the use of the statistical system *R* is gaining in popularity and is strongly endorsed by prominent computational linguists (Baayen 2008, Gries 2009). The downside of *R* is the steep learning curve and the dreaded command line interface (cf. Garretson 2008: 80).

What is crucial in visualization tools such as *Mondrian* and our *Text Variation Explorer* from the linguist's standpoint is the chance for rapid and interactive exploration of data. It is known that interaction enhances discovery, and linguistic data visualizations are no exception. Generating "Aha! That's interesting!" exclamations may succeed with static data visualizations, but the chances are far better with interactive visualization tools.

The challenge we now have is how to make visual analysis tools more accessible to the linguistic community. We believe that our *Text Variation Explorer* is a good trade-off between usability and utility, and serves as an example of a class of applications we need to strive for.

Notes

* This research was funded by the Academy of Finland (grant numbers 129300, 129350) and by Langnet, the Finnish graduate school in language studies.

References

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Connor, U. & Upton, T. A. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Binongo, J. N. G. 2003. "Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution". *Chance*, 16 (2), 9–17.

Brown Corpus. 1964, 1971, 1979. *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown)*. Compiled by W. N. Francis & H. Kučera. Providence, RI: Brown University.

Card, S. K., Mackinlay, J. D. & Shneiderman, B. 1999. *Readings in Information Visualization: Using Vision to Think*. San Francisco, CA: Morgan Kaufmann.

Craft, B. & Cairns, P. 2005. "Beyond guidelines: What can we learn from the Visual Information Seeking Mantra?" In *IV'05: 9th Annual International Conference on Information Visualisation*. Los Alamitos, CA: IEEE Computer Society, 110–118.

Culy, C. & Lyding, V. 2010. "Double Tree: An advanced KWIC visualization for expert users". In *IV 2010: 14th International Conference on Information Visualization*. Los Alamitos, CA: IEEE Computer Society, 98–103.

Garretson, G. 2008. “Desiderata for linguistic software design”. *International Journal of English Studies*, 8 (1), 67–94.

Gries, S. T. 2009. *Quantitative Corpus Linguistics with R*. New York: Routledge.

Hakuta, K. 2011. *WordSift: Supporting Instruction and Learning through Technology in San Francisco*. Washington, DC: The Council of the Great City Schools.

Hearst, M. 2009. *Search User Interfaces.* Cambridge: Cambridge University Press.

Hilpert, M. 2011. “Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora”. *International Journal of Corpus Linguistics*, 16 (4), 435–446.

ICE Corpus. 2014. *The International Corpus of English*. http://ice-corpora.net/ice/index.htm (accessed February 2014).

Kaufer, D., Geisler, C., Vlachos, P. & Ishizaki, S. 2006. “Mining textual knowledge for writing education and research: The DocuScope project”. In L. V. Waes, M. Leijten & C. M. Neuwirth (Eds.), *Writing and Digital Media*. Amsterdam: Elsevier, 115–129.

Keim, D. A. & Oelke , D. 2007. “Literature fingerprinting: A new method for visual literary analysis”. In W. Ribarsky & J. Dill (Eds.), *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology 2007, October 30 – November 1, Sacramento, CA, USA*. Piscataway, NJ: IEEE, 115–122.

Kilgarriff, A. 2012. “Review of M. Paquot (2010), *Academic Vocabulary in Learner Writing: From Extraction to Analysis*”. *International Journal of Corpus Linguistics*, 17 (1), 125–130.

Lijffijt, J., Papapetrou, P. & Puolamäki, K. 2012. “Size matters: Finding the most informative set of window lengths”. In P. A. Flach, T. De Bie & N. Christianini (Eds.), *Proceedings of the European Conference of Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML-PKDD 2012)*, Part II. Berlin and Heidelberg: Springer, 451–466.

LOB Corpus. 1970–1978. *The LOB Corpus, original version*. Compiled by G. Leech, Lancaster University, S. Johansson, University of Oslo (project leaders) & K. Hofland, University of Bergen (head of computing).

Pike, W. A., Stasko, J., Chang, R. & O’Connell, T. A. 2009. “The science of interaction”. *Information Visualization*, 8 (4), 263–274.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org (accessed March 2012).

Shneiderman, B. 1996. “The eyes have it: A task by data type taxonomy for information visualizations”. In *VL’96: Proceedings of the 1996 IEEE Symposium on Visual Languages*. Los Alamitos, CA: IEEE Computer Society, 336–343.

Siirtola, H. 2012. *Text Variation Explorer* (TVE). Available at: http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve.html (accessed April 2012).

Siirtola, H. 2013. TVE Video Tutorial. Available at: http://www.uta.fi/sis/tauchi/virg/projects/dammoc/tve/TVE.mp4 (accessed February 2013).

Siirtola, H., Nevalainen, T., Säily, T. & Räihä, K.-J. 2011. “Visualisation of text corpora: A case study of the PCEEC”. In T. Nevalainen & S. M. Fitzmaurice (Eds.), *How to Deal with Data: Problems and Approaches to the Investigation of the English Language over Time and Space*. Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/07/siirtola_et_al/ (accessed February 2013).

Spence, R. 2007. *Information Visualization: Design for Interaction*. Harlow: Prentice-Hall Europe, Pearson Education Ltd.

Theus, M. 2011. *Mondrian – Interactive Statistical Data Visualization in Java.* http://stats.math.uni-augsburg.de/mondrian/ (accessed February 2014).

Theus, M. & Urbanek, S. 2008. *Interactive Graphics for Data Analysis: Principles and Examples*. Boca Raton, FL: Chapman & Hall/CRC.

Ware, C. 2004. *Information Visualization: Perception for Design*. Second edn. San Francisco, CA: Morgan Kaufmann.
Youmans, G. 1991. "A new tool for discourse analysis: The vocabulary-management profile". *Language*, 67 (4), 763–789.

Authors' addresses

Harri Siirtola
School of Information Sciences
University of Tampere
Kanslerinrinne 1
FI-33014 University of Tampere
Finland

harri.siirtola@sis.uta.fi

Tanja Säily
Department of Modern Languages
University of Helsinki
P.O. Box 24 (Unioninkatu 40)
FI-00014 University of Helsinki
Finland

tanja.saily@helsinki.fi

Terttu Nevalainen
Department of Modern Languages
University of Helsinki
P.O. Box 24 (Unioninkatu 40)
FI-00014 University of Helsinki
Finland

terttu.nevalainen@helsinki.fi

Kari-Jouko Räihä
School of Information Sciences
University of Tampere
Kanslerinrinne 1
FI-33014 University of Tampere
Finland

kari-jouko.raiha@uta.fi